APPLICATION FOR
UNITED STATES PATENT
IN THE NAME OF

KEVIN B. STANTON

FOR

APPARATUS AND METHOD TO PROVIDE NODE-TO-NODE CONNECTIVITY WITH A SHARED REMOTE NETWORK INTERFACE DEVICE

Prepared By:

PILLSBURY WINTHROP LLP
725 South Figueroa Street, Suite 2800
Los Angeles, CA 90017-5406
Telephone (213) 488-7100
Facsimile (213) 629-1033

Attorney Docket No.: 81674-249725

Client Docket No.: P12817

Express Mail No.: EL 860 912 797 US

PATENT
81674-249725

TITLE OF THE INVENTION

APPARATUS AND METHOD TO PROVIDE NODE-TO-NODE CONNECTIVITY WITH A SHARED REMOTE NETWORK INTERFACE DEVICE

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention generally relates to a network interface to provide access to an Ethernet networking system. More particularly, the present invention relates to a remote virtual network interface that provides Ethernet connectivity to multiple InfiniBand nodes.

2. Discussion of the Related Art

Nodes, such as personal computers and engineering workstations, are conventionally interconnected to form local area networks ("LANs") that allow messages to be sent and programs to be downloaded, for example, from file servers on the LAN. Ethernet is a shared-media network architecture, defined in the Institute of Electrical and Electronics Engineers ("IEEE") 802.3 standard, republished October 16, 2000, and is currently the most widely used architecture for LANs. Ethernet uses both bus and star topologies, in which nodes are attached to a trunk segment, which is the primary piece of cable in an Ethernet network.

In a star configuration, several nodes are interconnected through a common hub or concentrator. A hub serves as a common termination point for multiple nodes and relays signals along the appropriate paths. Generally, the hub is a unit, having a number of connectors to which nodes are attached. Hubs usually accommodate a plurality of nodes (e.g., 4, 8, 12, 24, or more nodes), and many hubs include connectors for linking to other hubs. Each node in the network is typically a computer of some type, such as a personal computer ("PC"), Macintosh, minicomputer, or mainframe, where the computer generally includes a network interface card ("NIC") for interfacing the node to the hub to enable networking capabilities. In other words, each NIC generally interfaces only one node.

InfiniBand is a shared-media network architecture, developed to manage the increased traffic placed on LANs. InfiniBand is used to interconnect processor nodes and I/O nodes, forming a system area network ("SAN") that functions independently of the host operating system ("OS") and processor platform. InfiniBand is a point-to-point, switched I/O fabric that interconnects end node devices by cascaded switch devices. The InfiniBand Trade Association's ("IBTA") specification 1.0.a, republished June 19, 2001, defines the InfiniBand architecture, which offers greater bandwidth, increased scalability, and decreased CPU utilization, as compared to Ethernet. The IBTA projects that the bandwidth capacity of InfiniBand will remain superior to the bandwidth capacity of Ethernet by a factor of ten. However, when a NIC is used to provide Ethernet connectivity to multiple InfiniBand nodes, problems may arise if an InfiniBand node attempts to deliver data through the NIC to another InfiniBand node.

For example, when a packet is received by the NIC from an Ethernet node, the NIC determines to which node the data is destined and delivers the data to the destination node. When a packet is received by the NIC from an InfiniBand node, the NIC data is typically destined for an Ethernet node. However, if the packet is not destined for an Ethernet node, but rather is destined for another InfiniBand node (which may be on the same Internet Protocol ("IP") subnet), the Ethernet switch may fail to deliver the packet to the destined InfiniBand node, assuming that the destined InfiniBand node has already received the packet because the data came from a link in the direction of the destined InfiniBand node.

One solution may be to load a separate driver to implement an intra-InfiniBand LAN network emulation. However, this technique requires that the binding order/precedence from hostname to IP address ensures that traffic between the two InfiniBand nodes follows the intra-InfiniBand LAN emulation route, rather than the NIC route.

Another solution may be to disallow communication between the InfiniBand nodes on the subnet associated with the Ethernet port that is connected to the InfiniBand nodes. However, this technique eliminates the possibility of achieving the original goal of transferring information from one InfiniBand node to another.

Thus, a network interface that is capable of routing a data packet from one InfiniBand node to another InfiniBand node is required.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates a remote virtual network interface according to an embodiment of the present invention;

Fig. 2 illustrates a network system according to an embodiment of the present invention; and

Fig. 3 illustrates a flow chart for a method of routing a data packet from a first InfiniBand node to a second InfiniBand node according to an embodiment of the present invention.

DETAILED DESCRIPTION

Reference in the specification to "one embodiment", "an embodiment", or “another embodiment” of the present invention means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrase "in one embodiment" or "according to an embodiment" appearing in various places throughout the specification are not necessarily all referring to the same embodiment. Likewise, appearances of the phrase “in another embodiment” or “according to another embodiment” appearing in various places throughout the specification are not necessarily referring to different embodiments.

Fig. 1 illustrates a remote virtual network interface according to an embodiment of the present invention. The remote virtual network interface 100 includes an Ethernet receiving element 110, an Ethernet transmitting element 120, an InfiniBand receiving element 130, an InfiniBand transmitting element 140, a detector 150, and a routing element 160. The Ethernet receiving element 110 is in communication with an Ethernet node 170. The Ethernet transmitting element 120 is also in communication with the Ethernet node 170. The InfiniBand receiving element 130 receives a data packet from a first InfiniBand node 180. The data packet includes a destination indicator. The detector 150 reads the destination indicator and compares the destination indicator to a known value. The routing element 160 delivers the data packet from the InfiniBand receiving element 130 to the InfiniBand transmitting element 140. The InfiniBand transmitting element 140 transmits the data packet from the first InfiniBand node 180 to a second InfiniBand node 190.

According to an embodiment of the present invention, the destination indicator may be a destination media access control ("MAC") address. The known value may be a range of MAC addresses, where a range of MAC addresses is defined to be one or more MAC addresses. In one embodiment, the detector 150 and the routing element 160 may be within a single device. The remote virtual network interface 100 may be virtualized by implementing microcode in a network processor and/or a set of integrated circuits. A set of integrated circuits is defined as one or more integrated circuits.

Fig. 2 illustrates a network system according to an embodiment of the present invention. The network system 200 includes an Ethernet node 170, an Ethernet switch 210, a first InfiniBand node 180, a second InfiniBand node 190, an InfiniBand switch 220, and a remote virtual network interface 100. The Ethernet node 170 may receive a first data packet from the remote virtual network interface 100. The Ethernet switch 210 may select the Ethernet node 170 to receive a second data packet. The first InfiniBand node 180 may transmit a data packet to the remote virtual network interface 100. The data packet includes a destination indicator. The InfiniBand switch 220 may select the second InfiniBand node 190 to receive the data packet from the first InfiniBand node 180.

According to an embodiment of the present invention, the first data packet and the second data packet are same.

Fig. 3 illustrates a flow chart for a method of routing a data packet from a first InfiniBand node to a second InfiniBand node according to an embodiment of the present invention. Within the method and referring to Fig. 1 and Fig. 2, Ethernet connectivity is provided 310 to the first InfiniBand node 180 and to the second InfiniBand node 190. A remote virtual network interface 100 may receive 320 a data packet from the first InfiniBand node 180. The data packet includes a destination indicator. The detector 150 may read 330 the destination indicator. The destination indicator may indicate 340 that the data packet is to be delivered to the second InfiniBand node 190 by comparing the destination indicator to a known value. If the data packet is to be delivered to the second InfiniBand node 190, then the routing element 160 may deliver 350 the data packet to the second InfiniBand node 190. If the data packet is not to be delivered to the second InfiniBand node 190, then the data packet may be delivered 360 to the Ethernet node 170.

According to an embodiment of the present invention, the destination indicator may be a destination MAC address. The known value may be a range of MAC addresses, where a range of MAC addresses is defined to be one or more MAC addresses. The method of routing the data packet from the first InfiniBand node 180 to the second InfiniBand node 190 may include virtualizing the remote virtual network interface 100 by implementing microcode in a network processor and/or a set of integrated circuits. A set of integrated circuits is defined as one or more integrated circuits.

In short, the remote virtual network interface 100 according to the present invention provides Ethernet connectivity to multiple InfiniBand nodes. Specifically, the remote virtual network interface 100 is capable of routing a data packet from a first InfiniBand node 180 to a second InfiniBand node 190, even if the first InfiniBand node 180 and the second InfiniBand node 190 are on the same subnet. Furthermore, communication is allowed between the first InfiniBand node 180 and the second

InfiniBand node 190 regardless of whether an intra-InfiniBand LAN network is emulated, and regardless of the order and priority of binding.

While the description above refers to particular embodiments of the present invention, it will be understood that many modifications may be made without departing from the spirit thereof. The accompanying claims are intended to cover such modifications as would fall within the true scope and spirit of the present invention. The presently disclosed embodiments are therefore to be considered in all respects as illustrative and not restrictive, the scope of the invention being indicated by the appended claims, rather than the foregoing description, and all changes that come within the meaning and range of equivalency of the claims are therefore intended to be embraced therein.